Clustering the administrative districts for the city of São Paulo (SP, Brazil): A gastronomic adventure in the times of social isolation

Francisco Martellini

May 5, 2020

1. Introduction

This the report for the applied capstone project for IBM Data Science professional certificate at Coursera. The main objective from this exercise is use the concepts learned in all courses from certificate to solve a basic problem: compare different neighborhoods using the data from the Foursquare API.

The first approach for this this project, consist in use the data of the commercial locations from the São Paulo city (SP, Brazil) and the data from the COVID-19 pandemic available until this moment for each administrative district (neighborhood) of the city. However, the Brazil stay in a political crises in this moment, part because of the economic effects of the pandemic. In a ethical perspective, I can not create a representative analysis with this sources, and ensure that this results not will be used by the political powers in conflict if this report is open available in the internet, specially in this moment.

The solution was think this capstone more as a exploratory research than an heavy statistical analysis, inspired by the Japanese TV series, Samurai Gourmet. In that show, the character walk around the city searching for new restaurants in the improbable places. If the samurai wanted made the same in times of social isolation (or lockdown in certain regions of the world), he face a problem: how he can wanderer by the city if he need stay in home?

This exercise have another dimension too, because the samurai from the TV show is an elderly what makes he part of one of the risk groups for the COVID-19. To help our samurai in your honorable mission to know new flavors, this report use the cluster analysis in the administrative districts of the city of São Paulo, so he can walk and stay in home, using the delivery services. Who knows what he can find?

2. Data acquistion and cleaning

The data sources used was two: the information available for the city of São Paulo in the Foursquare and the dataset available in Kaggle with the geographic coordinates if the city and your administrative divisions. There was no need of cleaning data, but the Kaggle dataset was confronted with the city law that define the administrative division. This analysis was necessary to ensure that all 96 administrative districts is available in the dataset, in conformity with the most recent legislation.

The Foursquare Places API has a limit of 950 regular API Calls and return only 100 venues, but it was enough for didatic purposes. The dataset from Kaggle was available in a JSON file that can easily parsed in a Jupyter Notebook, by Caio B. Silva with a Creative Commons 1.0 Public Domain Dedication that allows copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission. The dataset containing each district's name as well as its latitude, longitude, and population. For the use in this activaty, the population data was not used, only the district name and the geolocations.

3. Exploratory data analysis

The Jupyter Notebook that was used to explore data is available in the my github repository for the Data Science Applied Capstone. The first step was to compare the Kaggle dataset with the City Law 11.220/1992 that define the geographic division from the area of the city of São Paulo in districts (this law replaced the previous definitions from the law 10.932/1991). According with this document, there are 96 districts in the city:

> Água Rasa, Alto de Pinheiros, Anhanguera, Aricanduva, Arthur Alvim, Barra Funda, Bela Vista, Belém, Bom Retiro, Brás, Brasilândia, Butantã, Cachoeirinha, Cambuci, Campo Belo, Campo Grande, Campo Limpo, Cangaíba, Capão Redondo, Carrão, Casa Verde, Cidade Adernar, Cidade Dutra,Cidade Líder, Cidade Tiradentes, Consolação, Cursino, Ermelino Matarazzo, Freguesia do Ó, Grajaú, Guaianases, Iguatemi, Ipiranga, Itaim Bibi, Itaim Paulista, Itaquera,Jabaquara, Jaçanã, Jaguara, Jaguaré, Jaraguá, Jardim Anqela, Jardim Helena, Jardim Paulista, Jardim Sao Luis, José Bonifácio, Lajeado, Lapa, Liberdade, Limão, Handaqui, Marsilac, Moema, Mooca, Morumbi, Parelheiros, Pari, Parque do Carmo, Pedreira, Penha, Perdizes, Perus, Pinheiros, Pirituba, Ponte Rasa, Raposo Tavares, República, Rio Pequeno, Sacomã, Santa Cecília, Santana, Santo Amaro, São Domingos, São Lucas, São Mateus, São Miguel, São Rafael, Sapopemba, Saúde, Sé, Socorro, Tatuapé, Tremembé, Tucuruvi, Vila Andrade, Vila Curuçã, Vila Formosa, Vila Guilherme, Vila Jacuí, Vila Leopoldina, Vila Maria, Vila Mariana, Vila Matilde, Vila Medeiros, Vila Prudente and Vila Sônia.

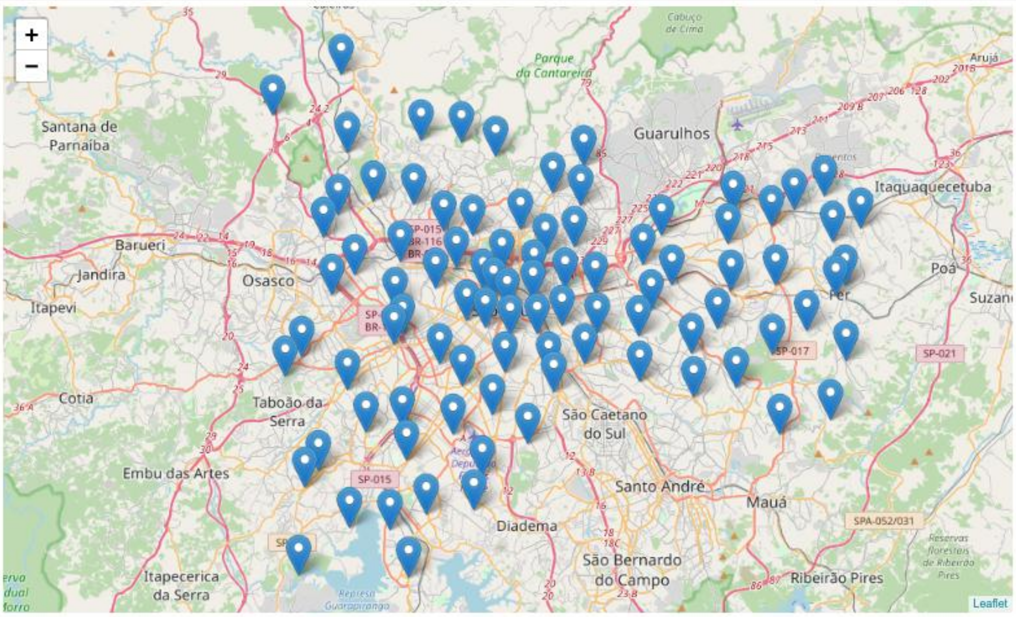
All of them are available in the dataset with the respective geographic coordinates in conformity with the limits defined by the law. To provide a visual verification a folium map was used (Picture 1), with the coordinates to the city, that are latitude -23.5506507 and longitude -46.6333824, founded with the Nominatim library. This analysis provide a visual approach that despite your imprecision, allows to ensure that all point is set in the city area.

Picture 1: Map of districts for the city of São Paulo and the python code.

```python
# create map of Sao Paulo using latitude and longitude values
map_saopaulo = folium.Map(location=[latitude, longitude], zoom_start=11)

# add markers to map
for lat, lng, district_name, population in zip(spdist_df['Latitude'], spdist_df['Longitude'], spdist_df['Distric
    label = 'Name: {}, Population: {}'.format(district_name,population)
    label = folium.Popup(label, parse_html=True)
    folium.Marker(
        [lat, lng],
        popup=label).add_to(map_saopaulo)

map_saopaulo
```



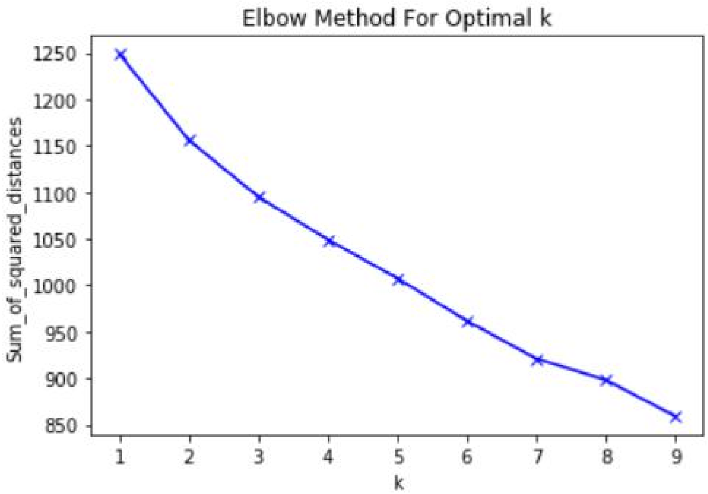
The Foursquare data analysis was made with the top ten venues for every district and despite the objective is made a analysis with the restaurants, we use all venues founded for every district to try evaluate if the other commercial points affect the restaurants.

4. Predictive modelling

In this project we use a classification model to analyse the data with cluster analysis, that try to group a set of objects in such a way that this objects are in the same group (cluster), or in a qualitative approach, they are more similar to each other than to those in other groups. The method used was K-Means clustering, but before to use this method it is necessary to find the optimal value of K (number of clusters) to apply the method.

The first approach to find the value of K is called Elbow Criterion. The idea behind elbow method is to run k-means clustering on a given dataset for a range of values of K, and for each value of k, calculate the sum of squared errors (SSE). After that, plot a line graph of the SSE for each value of K. If the line graph looks like an angle in the graph, the "elbow" on the arm is the value of optimal K (the number of clusters). Here, we want to minimize SSE. SSE tends to decrease toward 0 as we increase K and SSE is 0 when K is equal to the number of data points in the dataset, because then each data point is its own cluster, and there is no error between it and the center of its cluster. So the goal is to choose a small value of K that still has a low SSE, and the elbow usually represents where we start to have diminishing returns by increasing K. The plot is the Picture 2, and the “elbow” maybe be K=2 or K=3.

Picture 2: Plot for the Elbow Method to K-Means Cluster.

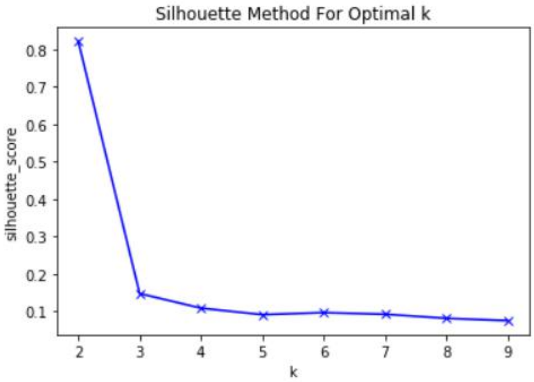


Elbow method does not seem to help us to determine the optimal number of clusters. So, we try to use another method: the Silhouette Method. This method measure how similar a point is to its own cluster (cohesion) compared to other clusters. A higher Silhouette Coefficient score relates to a model with better-defined clusters. The Silhouette Coefficient is defined for each sample and is composed of two scores: (i) The mean distance between a sample and all other points in the same class; and (ii) The mean distance between a sample and all other points in the next nearest cluster. To find the optimal value of K in this case, we need to loop through 1 to n for and calculate Silhouette Coefficient for each sample. A higher Silhouette Coefficient indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

The plot is in the Picture 3 with the Silhouette Coefficients for every value of n (with n from 1 to 10). The best value seems n=2 with coefficient 0,82, but n=3 have more similarity with the result find in the Elbow Method despite the coefficient was 0,14, so the best value is n=3.

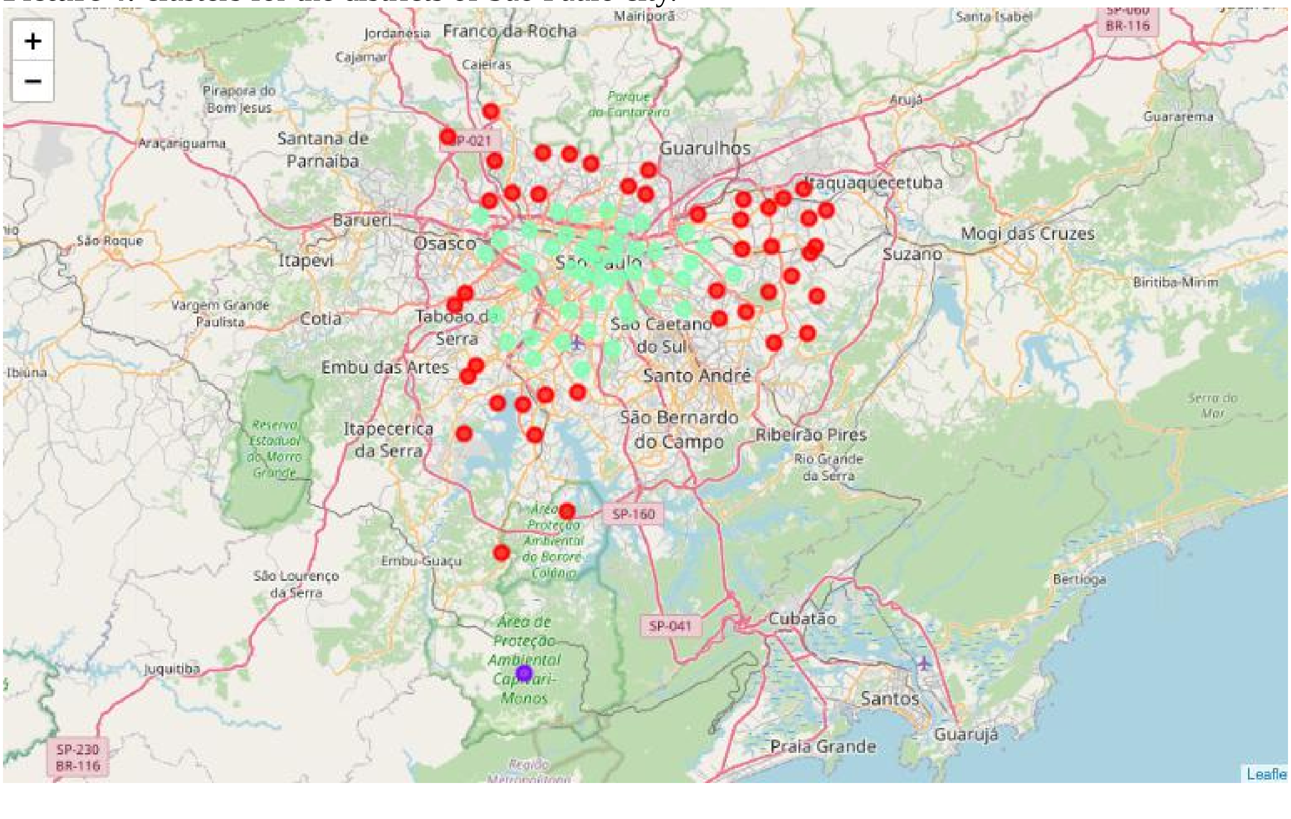
Picture 3: Plot for the Silhouette Method to K-Means Cluster.

```
For n_clusters=2, The Silhouette Coefficient is 0.8210573962604745
For n_clusters=3, The Silhouette Coefficient is 0.1475455275919334
For n_clusters=4, The Silhouette Coefficient is 0.10848908210594288
For n_clusters=5, The Silhouette Coefficient is 0.09116363507628773
For n_clusters=6, The Silhouette Coefficient is 0.09659645843672576
For n_clusters=7, The Silhouette Coefficient is 0.09220412097070717
For n_clusters=8, The Silhouette Coefficient is 0.0813724485647067
For n_clusters=9, The Silhouette Coefficient is 0.07520483576327115
```



Applying the K-Means with K=3 and visualizing the resulted clusters in the city map, we find the result of the Picture 4 (red is cluster 1, pink is cluster 2 and green is cluster 3). The cluster 2 is isolated in the corner of the map, because this district is located in a environmental protection area and have unique characteristics when compared with the great urbanization process of the city. The other two cluster let us to perceive the separation between the center and the suburbs of the city, that is characteristic from the urban development of the major cities in Brazil, and show how similar is the social and urban inequality in this districts.

Picture 4: clusters for the districts of São Paulo city.



Analyzing each cluster we find the most common venue for everyone.

Cluster 1		**Cluster 2**		**Cluster 3**	
Bakery	44	Brazilian Restaurant	1	Pizza Place	38
Pizza Place	41	Flower Shop	1	Bakery	32
Gym / Fitness Center	41	Flea Market	1	Pet Store	32
Brazilian Restaurant	40	Film Studio	1	Dessert Shop	31
Dessert Shop	30	Field	1	Ice Cream Shop	27
Gym	29	Food & Drink Shop	1	Italian Restaurant	27
Bar	25	Food	1	Burger Joint	24
Japanese Restaurant	20	Zoo	1	Brazilian Restaurant	22
Restaurant	17	Fish & Chips Shop	1	Gym / Fitness Center	22
Pet Store	14	Food Court	1	Bar	18

5. Conclusions and future approach

The results show us how our hypothetical samurai can try to find new restaurants and saty in home during the pandemic. In the future, we need to use better datasets for commercial points.